BMC Bioinformatics

BioMed Central

Methodology article

Open Access

GONOME: measuring correlations between GO terms and genomic positions

Stefan M Stanley, Timothy L Bailey and John S Mattick*

Address: Institute for Molecular Bioscience, University of Queensland, Brisbane 4072, Australia

Email: Stefan M Stanley - s.stanley@imb.uq.edu.au; Timothy L Bailey - t.bailey@imb.uq.edu.au; John S Mattick* - j.mattick@imb.uq.edu.au

* Corresponding author

Published: 25 February 2006

*BMC Bioinformatics*2006, **7**:94 doi:10.1186/1471-2105-7-94

Received: 06 December 2005

Accepted: 25 February 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/94

Abstract

Background: Current methods to find significantly under- and over-represented gene ontology (GO) terms in a set of genes consider the genes as equally probable "balls in a bag", as may be appropriate for transcripts in micro-array data. However, due to the varying length of genes and intergenic regions, that approach is inappropriate for deciding if any GO terms are correlated with a set of genomic positions.

Results: We present an algorithm – GONOME – that can determine which GO terms are significantly associated with a set of genomic positions given a genome annotated with (at least) the starts and ends of genes. We show that certain GO terms may appear to be significantly associated with a set of randomly chosen positions in the human genome if gene lengths are not considered, and that these same terms have been reported as significantly over-represented in a number of recent papers. This apparent over-representation disappears when gene lengths are considered, as GONOME does. For example, we show that, when gene length is taken into account, the term "development" is not significantly enriched in genes associated with human CpG islands, in contradiction to a previous report. We further demonstrate the efficacy of GONOME by showing that occurrences of the proteosome-associated control element (PACE) upstream activating sequence in the *S. cerevisiae* genome associate significantly to appropriate GO terms. An extension of this approach yields a whole-genome motif discovery algorithm that allows identification of many other promoter sequences linked to different types of genes, including a large group of previously unknown motifs significantly associated with the terms 'translation' and 'translational elongation'.

Conclusion:GONOME is an algorithm that correctly extracts over-represented GO terms from a set of genomic positions. By explicitly considering gene size, GONOME avoids a systematic bias toward GO terms linked to large genes. Inappropriate use of existing algorithms that do not take gene size into account has led to erroneous or suspect conclusions. Reciprocally GONOME may be used to identify new features in genomes that are significantly associated with particular categories of genes.

Background

The Gene Ontology (GO) project [1] arose partly in response to the problem of non-uniform assignment of genomic annotations. Biological databases are notorious for the inconsistency of their annotation terminology, and attempts to apply statistical methods based on annota-

tions across, or even within, genomes face difficulty with this problem. GO addresses this issue by re-expressing annotations using controlled vocabularies, or ontologies; by providing a mechanism for formalizing relationships between qualitative properties (GO terms) that can be associated to genomic features; and by creating a hierarchical structure of these qualities through the use of 'is-a' and 'part-of-a' relationships, allowing one to fit all annotation terms into a "tree" structure (actually a directed acyclic graph) with the most general terms at the root and the most specific terms as leaves. The GO database is broken into three such hierarchies, or aspects: biological process, molecular function and cellular component.

GO has become a popular way of analyzing sets of genes to find under- or over-represented terms associated with that set of genes, especially in expression micro-array datasets. One may, for example, apply a "GO analysis" to sets of up- or down-regulated genes to assess which processes or functions are undergoing coordinated regulation. A variety of web-based tools exist that allow one to enter a list of gene identifications and find the over- and under-represented GO terms associated to those genes – for example GOstat [2] and GO::TermFinder [3].

In this work, we consider the slightly different task of determining whether a *set of genomic positions* is associated with any GO terms. This situation arises in many contexts (see, e.g. [4,5]). One might, for example, wish to determine if a particular regulatory sequence motif is significantly associated with genes involved in any particular pathway.

The typical "GO analysis" begins with a set of genes and uses a random model that assumes that each gene in the genome is equally likely *a priori* to be included in the set. This assumption is inappropriate when the input is a set of genomic positions rather than a set of genes. Due to the varying length of genes and intergenic distances, randomly selected genomic positions are much more likely to fall within large genes or within large adjacent intergenic regions. Therefore, when determining which (if any) GO terms are significantly associated with a set of genomic positions, a different random model is required.

As an illustration, imagine that we are given a set of genomic positions and asked to determine with which GO terms they are associated. Note, firstly, that the GO database maps genes to terms, so we must define how we are going to map genomic positions to genes. A natural way to associate GO terms with genomic positions is to associate each position with a single gene, and, transitively, with that gene's GO terms: *position→gene→GO Term*. In this example, we map each (strand-specific) genome position falling within a gene to that gene. Suppose then that the genome consists of five 1 Kb genes annotated as metabolic and one 1 Mb gene annotated as meiotic. A randomly chosen genomic sequence in this genome is $10^6/5000 = 200$ times more likely to lie within the meiotic gene than within any of the metabolic genes. Therefore, a random model that assumes all genes (and, hence, their GO terms) are equally likely to be selected is clearly inappropriate. Using such a model would cause randomly chosen genomic positions to (erroneously) correlate with GO terms associated with the meiotic gene. Thus a new approach is required to assess the statistical significance of GO terms associated to genomic positions that explicitly considers gene length, as opposed to the event-based associations currently used with gene expression data.

Results

GONOME: Gene Ontology correlations in the genome

We have developed a new application, called GONOME [6], which calculates the statistical significance of the correlation between a set of genomic positions and their associated Gene Ontology terms. GONOME does this by applying a random model that assumes that each *position* (rather than each *gene*) in the portion of the genome under consideration is equally likely. This implies that the chance of a uniformly distributed random position "hitting" (lying within, or adjacent to) a particular genomic region is proportional to the size of the region, removing the bias toward large genes caused by considering all genes equally probable.

GONOME takes as input a set of genomic positions and a genome annotated with the locations of genes. The positions in the input set may be strand-specific or not. If a DNA strand is not specified, GONOME replaces the position with two positions: one on each DNA strand.

GONOME also allows the user to define which genomic regions are of interest in a particular analysis. The upstream, transcribed and downstream regions of each gene may be associated with its GO terms or treated as 'unscored'. These regions are linked to the GO terms associated with the gene's GO terms (If a gene has no associated GO terms, its regions are associated with the "placeholder" term "NO_GO"). When the upstream and downstream regions of two adjacent genes overlap, GONOME treats the positions in the overlap as lying in both regions. All positions not lying in the upstream, transcribed or downstream region of any gene are also considered unscored, have no GO terms associated with them, and the user can choose to either include or exclude them from the analysis.

The user can also control the allowed size of the associated upstream and downstream regions via configurable "cut-

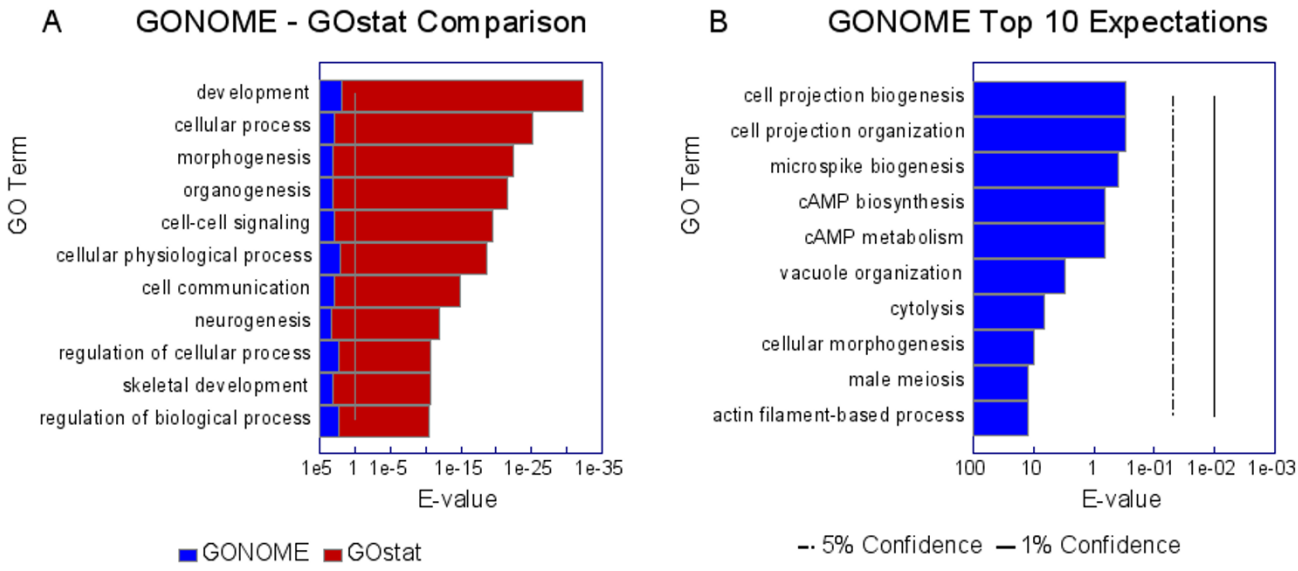

Figure 1
GONOME and GOstat output on random positions. Panel **A** compares GONOME and GOstat analyses of 30,000 randomly selected positions in the human genome. The *E*-values of the top 10 over-represented GO Terms as found by GOstat (red), and the values GONOME derives for the same terms (blue). Panel **B** shows the top ten over-represented terms according to GONOME. *E*-values were calculated as described in Methods.

offs". This is a useful feature because which genomic positions are of interest depends on the organism being studied as well as the type of positions being analyzed. For example, a position 100 bp upstream of a gene might naturally be associated with that gene, a position 500 bp upstream of one gene and 500 bp downstream of another might be ascribed to either or both of the flanking genes, while one 10 Kbp upstream and downstream from the nearest genes might not be associated at all. Such judgments are partly dependent on the size of the genome and the intergenic regions (e.g. yeast vs. human). Additionally, if the positions in the input set are (putative) promoters, one might only be interested in the region upstream of the gene, and only if within 500 bp.

Finally, rather than merely counting the number of times each GO term is associated with a genomic position in the input set, GONOME allows the user to specify weights for each type of region: upstream, transcribed and downstream. This allows the user to tailor a "correlation scoring function" appropriate to the biological questions being asked.

Over-represented terms from random positions

To validate the approach and compare it with previous approaches, we compared the statistical significance of the GO associations of a set of 30,000 randomly chosen strand-specific positions in the human genome calculated by GONOME and GOstat [2]. The genomic positions were generated uniformly across both strands of the genome. We set the GONOME parameters so that all three types of region (upstream, downstream and the transcribed gene, including its introns and UTRs) were included in the analysis and that all intergenic regions were ascribed as the downstream and upstream regions of the respective adjacent genes. For input to GOstat, a set of genes corresponding to the randomly chosen set of genomic positions was constructed similarly by adding the gene for positions falling within the boundaries of the transcribed gene, and by adding both the upstream and downstream genes for positions falling within intergenic sequences. The results are shown in Figure 1.

It is apparent from Figure 1A that GOstat reports numerous GO terms as being over-represented among the random positions. The most over-represented term is 'development', with the estimated probability of random occurrence (*E*-value) being 2.41×10^{-31} – a number that would be hard to consider insignificant. This result may be understood by observing that many of the over-represented terms are associated with genes encoding developmental regulators and/or membrane proteins, genes longer than most – the average length of all GO annotated genes is 106.5 Kb (5257 genes) (note that non-GO annotated genes have an average length of 34.4 Kb, suggesting that these may not include distal exons and/or include artifacts), whereas the average of those annotated as development is 133.9 Kb (770 genes), cellular process 125.8 Kb

Table 1: Novel putative motifs found by GONOME. GONOME was used to find over-represented GO terms associated with each possible n-mer (for *n* from 5 to 11) in the *S. Cerevisiae* genome, here are some of the significant motifs not reported elsewhere. (Motifs are expressed in IUPAC extended DNA alphabet: K is G or T; V is G, C or A; B is G or T or C; S is G or C; M is A or C)

Motif	GO Term	GONOME *E*-value	Transcription Factors
CCCCTAAAA	vitamin metabolism	2.5e-7	ADR1, NRG1
GCCCTAA	rRNA modification	1.1e-5	NRG1
TCCGCGG	Response to drug	8.7e-11	SUT1, STB5
GGVBCCSG	Translation	3.3e-30	-
CACGTGA	Sulfur amino acid metabolism	6.5e-10	CBF1
GKKGSMAAA	Protein catabolism	1.0e-10	
TGGCAAA	Protein catabolism	7.0e-4	-

(2832 genes), morphogenesis 152.2 Kb (466 genes), and organogenesis 158.3 Kb (392 genes). In contrast, GONOME finds no terms occurring significantly more often than expected at random (Figure 1A). What is more, the top ten GONOME *E*-values range from values of about one to about ten, as they should when the genomic positions are chosen at random. This result clearly demonstrates the inappropriateness of considering all genes as equally likely when analyzing genomic positions. This error has been made in a number of recent papers, which typically ascribe development as a significantly enriched term. For example, more than half of the over-represented terms reported in Table 1 from Seipel *et al*. [5] are also predicted as over-represented by GOstat applied to sets of random positions. Another example is provided by the Robinson *et al*. [7] analysis of CpG islands in the human genome, which we re-analyze in the next section. On the other hand, the general gene categories (RNA processing, regulation of transcription and development) reported as being significantly associated with ultraconserved elements [4] remain significantly associated using GONOME (data not shown).

Human CpG Islands

A CpG island is a cluster of CG dinucleotides, which appear as obvious features in mammalian genomes wherein only a quarter of the expected number of CG dinucleotides occur. CpG islands have previously been shown to be strongly associated with 'housekeeping' genes [8-12], but were recently reported to be also significantly associated with genes annotated with the GO term "development" using a chi-squared based method [7]. Following Robinson *et al*. [7], we applied GONOME to the positions of CpG islands in the human genome, restricting the scored positions to those occurring within genes and their 'promoter' regions (2000 bp upstream), against a null model that compares the observed pattern to that expected from random positions occurring uniformly throughout the entire genome (i.e. including the regions outside genes and their immediate 5'-promoter regions, termed the 'unscored' regions) (Figure 2).

The results show that, using the null model of the random incidence of CpG islands in the entire genome, the GO root node 'biological_process' and other broad functional descriptors are very significantly over-represented terms associated with CpG islands (Figure 2A). This occurs because CpG islands occur more often in or near the beginnings of genes *per se* than would be expected for a uniform distribution of random positions across the genome. Other generic terms indicating housekeeping functions as well as the term 'NO_GO' (the placeholder for genes without associated GO annotations) also receive significant *E*-values. Thus, if GONOME is applied against the whole genome in this way, and general GO terms are reported as being significantly over-represented, one may conclude that the chosen feature (in this case CpG islands) is strongly associated with genes and their promoters, genome-wide. However, while this may be generally informative, the strength of the signal obscures potential specific associations of the feature in question with particular subsets of genes, and thus one needs to be judicious about the choice of the null model.

The alternative null model excludes unscored positions in the genome, thereby calculating the chance that any GO term attains its score given the actual number of scoring positions. Using this model one can better assess whether there is a significant bias in the association of CpG islands with specific GO terms. When we repeated the analysis with this model, we found that GO terms with a "housekeeping" nature, such as those including the words 'metabolism,' 'transcription' and 'regulation' predominate (Figure 2B). This confirms the strong association of CpG islands with housekeeping genes, as well as the over-representation of the most significantly enriched term found in CpG island associated genes from Table 2 in Robinson *et al*. [7], 'regulation of transcription, DNA dependent'. However, of the remaining nine biological process terms in that table, seven are found to be over-represented in four runs of 30,000 *random* positions through GOstat. It should be borne in mind, however, that the previous analysis also used a tissue-specific metric on gene expression

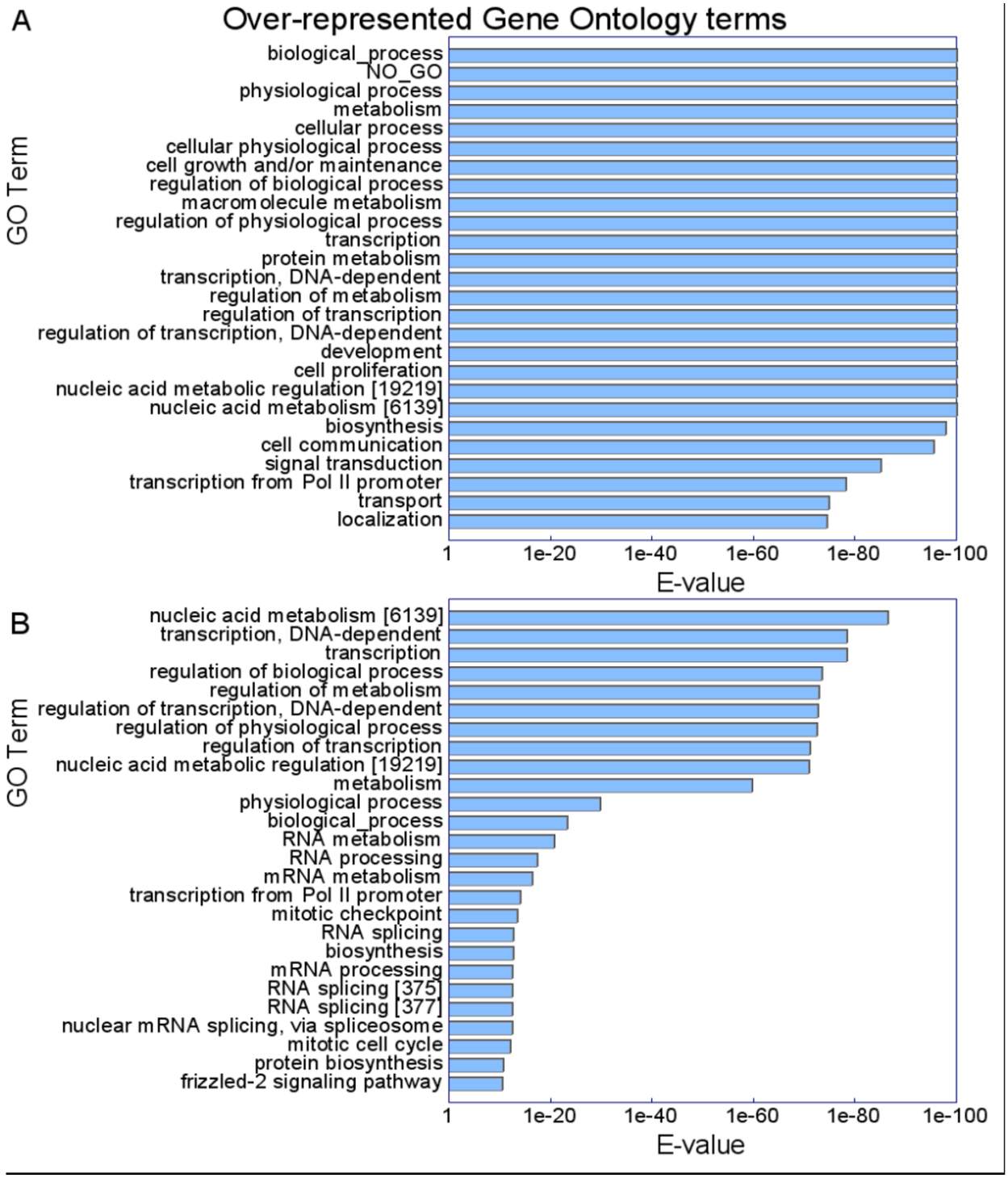

Figure 2

Over-represented gene ontology terms associated with human CpG islands. Over-represented gene ontology terms associated with human CpG islands as determined by GONOME when (**A**) unscored regions are included in the analysis and when (**B**) unscored regions are excluded from the analysis. The *E*-values of the 25 most over-represented GO process associated with CpG islands in the human genome in each case are shown. The image is the actual output of the GONOME application, save that long GO terms have been replaced with shorter equivalents and their GO identification numbers provided in brackets.

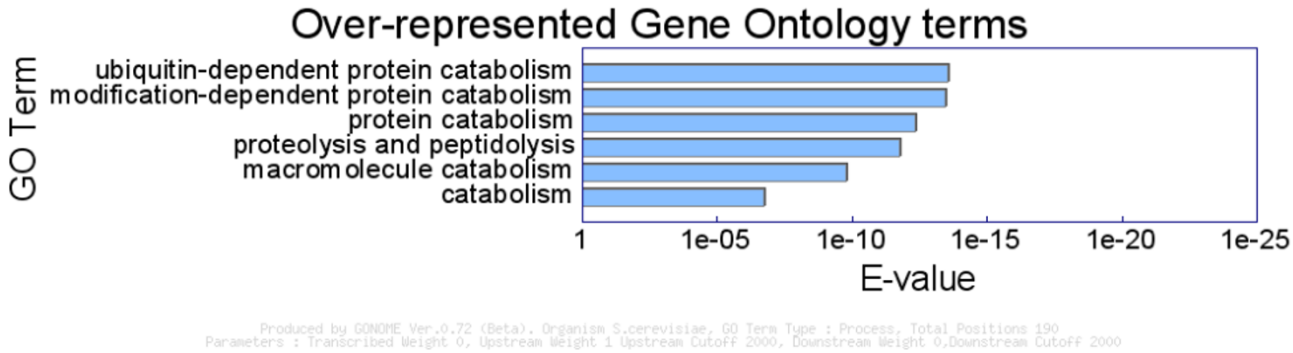

Figure 3
GONOME analysis of PACE elements. Figure legend text. Over-represented GO terms associated with positions in the *S. cerevisiae* genome of the proteosome associated control element (PACE) upstream activating sequence (UAS), 5'-GGT-GGCAAA-3'. "Locus" refers in general to a gene and its associated upstream and downstream regions, which are "Hit" once when a position falls within any associated region.

(the distribution of ESTs belonging to the same UniGene cluster) [7] that ours does not.

The dataset used in Figure 2 has been masked for repetitive sequences, and contains 67,697 positions. However, when GONOME is used with the same parameters on the unmasked set of 102,064 positions, virtually identical results are found. This demonstrates that GONOME is robust to noise – the repetitive positions showed no significantly over-represented terms (graphs available on the website). Full graphs and corresponding results for the murine and yeast genomes, as well as correlations with terms in the function and component GO hierarchies are available online [13].

Whole-genome motif discovery with GONOME

In another test of GONOME, we used the positions of all 190 exact occurrences of the proteosome associated control element (PACE) upstream activating sequence (UAS), 5'-GGTGGCAAA-3' [14] in the *S. cerevisiae* genome. The PACE sequence motif is known to be present in the upstream regions of 27 proteasomal genes, and a number of ubiquitin-proteasomal genes. As we are interested in a UAS, we directed GONOME to include only (2000 bp) upstream regions in the analysis, against a null model of the whole genome (the results of which barely differ from null model that only considers the scored regions, because highly gene-specific terms do not give a broader signal associated with genes generally). As expected, GONOME reveals significant over-representation of terms for the appropriate (proteasomal) processes (Figure 3) indicating its accuracy.

This analysis naturally raises the question of whether GONOME could be applied on a genome-wide scale to extract functionally-linked but as yet unrecognized UAS sequences that may bind other transcription factors. To this end, we used GONOME to analyze the positions in the yeast genome of all possible *k*-mers (for *k* in the range 5 to 11) looking for any that showed significant over-representation with particular GO terms. The complete results are available at [13], and largely recapitulate results of a similar study carried by Cora et al [15], without requiring the prior isolation of upstream sequences associated to groups of genes. Moreover, GONOME automatically identified several additional motifs, including the PACE element, not reported in Cora *et al*. (Table 1).

One interesting finding is that it would appear that the PACE element (GGTGGCAAA) is actually a specific subset of a more general motif (GKKGSMAAA) associated with most protein catabolism genes. Whereas the PACE element is found in the 2000 bp upstream of 27 genes annotated as involved in protein metabolism, with 18 of those also being annotated with 'protein catabolism', the related motif TGGCAAA (underlined above, a subset of both the PACE and the general motif) occurs in the upstream regions of 56 protein catabolism genes, and of 65 genes involved in 'macromolecule catabolism', perhaps implying greater promiscuity for the PACE associated transcription factor, *RPN4* [14,16,17], than is currently accepted, or the existence of another transcription factor with overlapping specificity.

Another noteworthy finding is that, along with the previously reported large groups of motifs associated with ribosome biogenesis and DNA replication, GONOME identified a large group of previously unknown motifs significantly associated with the terms 'translation' and 'translational elongation'. A set of genes is known to be regulated as a group, for example, in stress induced inhibition of translation [18]. The full set of over-represented

motifs found in this analysis can be also found in the resources section of the website.

Discussion

GONOME provides a flexible way to examine correlations between GO terms or other annotated features of genes and sets of genomic positions. Configurable parameters allow the genomic areas of interest to be defined and given relative weights. Two different null models allow the user to consider or exclude non-associated regions of the genome from the statistical analysis. GONOME reports correlations as *E*-values, conservatively accounting for the testing of multiple hypotheses. The statistical model accurately accounts for the varying length of genomic regions.

GONOME is also a useful adjunct to many positional genomic analysis methods (*e.g.*, BLAST [19] or linkage analysis) providing a generally applicable method for enhancing understanding of the biological significance of any set of genomic positions. It should also be noted that the methodology is not restricted to GO term data. In theory, any type of annotation data that can be associated with genes can be accommodated.

As with any statistical method, some caution needs to be used in the choice of parameters in order to achieve the best results. We have observed that, in large genomes, limiting the upstream and downstream cutoff distances to around 10 Kbp avoids penalizing genes next to gene deserts. It should also be borne in mind that larger genes may and probably do have more extended regions of regulatory information, including that in introns, and therefore it may be appropriate, depending on the context of the question, to (also) use incidence-based statistical packages such as GOStat to obtain another perspective on regulatory correlations.

Another issue can arise when multiple "hits" to a small number of genes occurs due to clustering. For example, if considering the terms associated to transposon positions, a heavily invaded gene with 100 Alu hits causes all terms associated with that gene to have significant *E*-values. While this is indeed a statistically significant result, it may obscure a more interesting genome-wide pattern. These types of issues can be handled by representing clusters of nearby positions with a single representative position. The GONOME software package includes a simple clustering routine for this purpose. The clustering routine finds all chains of positions separated by less than a (user-specified) threshold distance, and replaces them with the midpoint of the chain.

GONOME presently reports *E*-values computed by applying the conservative Bonferroni adjustment to *p*-values to correct for multiple hypotheses. In the future this might be extended to methods such as False Discovery Rate [20-22]. However, while the optimal way to account for multiple hypotheses in the densely inter-related gene ontology hierarchy remains an open question, the Bonferroni approach seems prudent.

Conclusion

GONOME provides a method for assessing the statistical significance of the association of genomic features with particular types of genes, and enables the correction of artifacts associated with variable gene size when using event-based statistical packages. GONOME may be tailored to specify the length of flanking sequences included in the analysis as well as used as a tool to discover new sequence motifs that are significantly associated with particular types of genes.

Methods

Correlations between GO terms and genomic positions

The objective of GONOME is to ascertain if a set of genomic positions is correlated with some biological property ("term") as annotated in the gene ontology (GO) database, and to compute the probability of such a correlation occurring at random. The GO database associates terms with genes, not genomic positions, so one first must define when a genomic position is considered to be associated with a GO term. A scoring function is then defined that measures the degree of correlation between a given go term, T, and a set of genomic positions, $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$. Finally, to determine if the degree of correlation is statistically significant, the random distribution of the scoring function is computed.

Associating GO terms with genomic positions

Each genomic position, x_i, in the input set, $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$, consists of a chromosome or contig identifier, a position on that chromosome or contig, and a strand (*i.e.*, Watson or Crick). (The user may specify unstranded positions in the input to GONOME, but these are each replaced by two positions, one on each DNA strand.) GONOME associates all the GO terms associated with a gene with each genomic position (on the gene's DNA strand) that falls within the upstream, transcribed or downstream region of the gene. GONOME permits the user to define the extents (in base pairs) of upstream and downstream regions. The downstream region of a gene extends until the start of the flanking gene's transcribed region or until the downstream cutoff is reached, whichever comes first. A similar definition applies to upstream regions. If the downstream region of a gene overlaps the upstream region of the flanking gene, the shared genomic positions (herein called an "overlap region") are associated with the GO terms of both genes. Genomic positions that are not part of any gene's upstream, downstream or

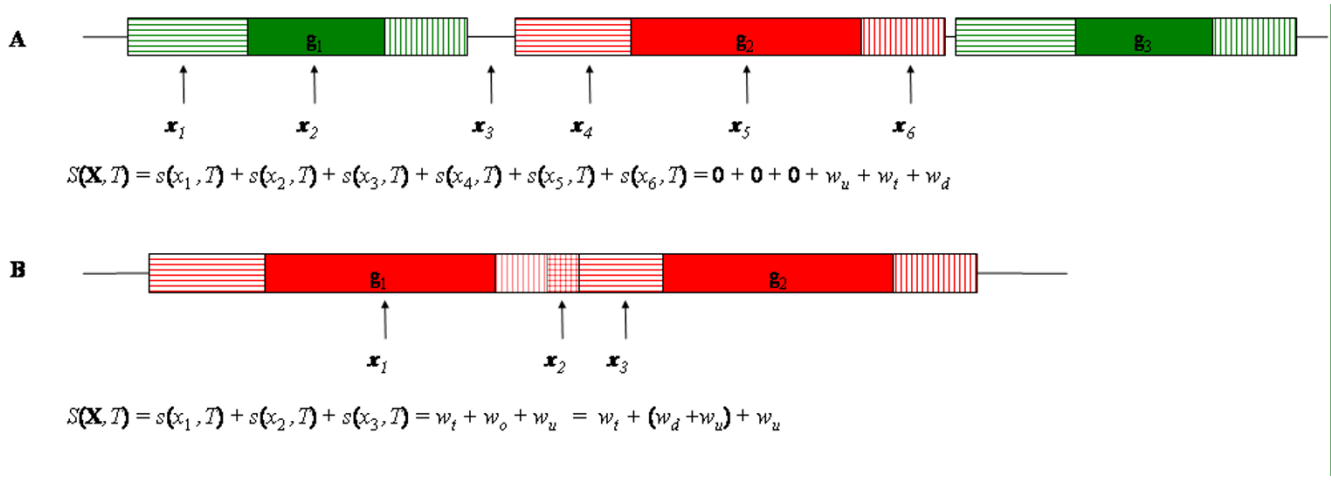

Figure 4
GONOME scoring function: *S*(X, *T*). Genomic positions in the input set X are shown with arrows. Regions belonging to genes annotated with GO term *T* are shaded red. Regions belonging to other genes are shaded green. Transcribed regions are shown in solid color. Upstream (downstream) regions are shown with horizontal (vertical) crosshatching. Unscored regions are shown as horizontal black lines. Positions in green and unscored regions receive association score of zero. Other positions receive association scores equal to the appropriate region weight. Panel **A** illustrates the simplest case where upstream and downstream regions of adjacent genes do not overlap. In panel **B**, the position x_2 lies in the "overlap" region of two "red" genes, so its score is the sum of the upstream and downstream weights.

transcribed regions (as defined above), are treated as unscored by GONOME and are not associated with any GO terms.

The correlation scoring function

GONOME assigns a "weight" to each type of genomic region that reflects the "strength" of the association between positions in the given type of region and the GO terms of the corresponding gene. The weights for the upstream, downstream, transcribed and overlap regions are called w_u, w_d, w_t and w_o, respectively. An upstream, transcribed or downstream region is defined to be "annotated with GO term *T*" if its associated gene is. An overlap region is defined to be annotated with *T* if *both* flanking genes are. Using these definitions, GONOME's score for the association between a single genomic position, *x*, and a GO term, *T*, in terms of the region weights is

$$s(x,T) = \begin{cases} w_u & \text{if } x \text{ is in an } \textit{upstream} \text{ region annotated with } T, \\ w_t & \text{if } x \text{ is in a } \textit{transcribed} \text{ region annotated with } T, \\ w_d & \text{if } x \text{ is in a } \textit{downstream} \text{ region annotated with } T, \\ w_o = w_u + w_d & \text{if } x \text{ is in an } \textit{overlap} \text{ region annotated with } T, \\ 0 & \text{otherwise.} \end{cases}$$

GONOME's scoring function for the association between the input set, **X**, and GO term, *T*, is the sum of the association scores of each genomic position in **X**,

$$S(\mathbf{X},T) = \sum_{x \in \mathbf{X}} s(x,T).$$

The GONOME correlation scoring function is illustrated in Figure 4.

The user may choose the value of each region weight in order to design a correlation scoring function appropriate to the task at hand. Weights must be non-negative. For example, to have the scoring function simply count the number of positions in the input set that lie within or near genes associated with GO term *T*, the GONOME user can set the region weights to $w_u = w_t = w_d = 1$. (This is not strictly true if any input positions lie in the overlap regions of two "red" genes as shown in Figure 4B). Such positions are effectively counted twice. On the other hand, if the user believes that positions in transcribed regions should be stronger evidence of association with term *T*, this intuition can be incorporated into the scoring function by setting the weights to, for example, $w_u = w_d = 1/2$ and $w_t = 1$. As a final, simpler, example, setting the weights to $w_u = 1$ and $w_u = w_d = 0$ will cause the scoring function to count only positions within upstream regions, as might be appropriate when studying promoter regions.

Determining statistical significance

Sufficiently large values of the scoring function $S(\mathbf{X},T)$ indicate that genomic positions annotated with GO term *T* are overrepresented in the set **X**. To determine how large is "large enough," the probability that $S(\mathbf{X},T) \geq S$ is estimated under the null assumption that the set of positions in **X** are chosen randomly from some "universe" of positions. This probability is commonly referred to as the *p*-

value of S. To adjust for testing multiple hypotheses, we convert p-values to E-values by multiplying by the number of terms in the GO ontology. This is also referred to as the "Bonferroni adjustment" [23] and gives a conservative estimate of the expected number of GO terms that would score S or more when the set of positions, **X**, is uncorrelated with any GO term.

The "universe" of positions referred to in the previous paragraph usually will be just the "genic" positions – the upstream, downstream, transcribed and overlap regions of genes. Alternatively, the user can specify that the universe include all positions in the genome.

GONOME computes score p-values by summing the (approximate) probabilities of all possible ways to achieve a score of S or greater. Each possible score is the sum of a total of n weights, one for each element in **X**. So, if variables u, t, d and o represent the number of positions in **X** that lie in upstream, transcribed, downstream and overlap regions, respectively, the correlation score is $S(\mathbf{X},T) = uw_u + tw_t + dw_d + ow_o$. Let $\Pr(n,u,t,d,o)$ be the probability that, out of n randomly chosen genomic positions, u, t, d and o, respectively are in upstream, transcribed, downstream and overlap regions. Then, by definition, the p-value of score S is

$$\Pr(S(\mathbf{X},T) \geq S) = \sum_{u=0}^{n} \sum_{t=0}^{n-u} \sum_{d=0}^{n-u-t} \sum_{o=\lceil (S-uw_u-tw_t-dw_d)/w_o \rceil}^{n-u-t-d} \Pr(n,u,t,d,o),$$

Equation 1

provided that $w_o \neq 0$. (If any region weights are zero, those regions are treated as unscored regions, and Equation 1 is modified to only sum over regions with non-zero weights.)

GONOME estimates $\Pr(n,u,t,d,o)$ in Equation 1 using the multinomial distribution

$$M(u,t,d,o,z,p_u,p_t,p_d,p_o,p_z) = \frac{n!}{\prod_{r\in\{u,t,d,o,z\}} r!} \prod_{r\in\{u,t,d,o,z\}} (p_r)^r,$$

Equation 2

where $z = n - u - t - d - o$ is the number of positions in **X** that lie in unscored regions, and p_r is the fraction of positions of type r in the "universe" of positions. Equation 2 will be a good estimate of $\Pr(n,u,t,d,o)$ as long as n is small relative to the size of the "universe," so that the probability of randomly choosing the same position twice as very small [24]. This is because Equation 2 represents the probability of randomly selecting the specified numbers (u, t, d, o and z) of different types of positions in n random selections *with replacement*, whereas the correct analogy should be selecting *without replacement*, since the input (**X**) is a set and a single position cannot appear (*i.e.*, be selected) more than once.

Optimizing the p*-value computation*

When all region weights are non-zero, computing Equation 1 requires $O(n^4)$ operations. This becomes prohibitive for large values of n. GONOME reduces the computational requirements by a factor of n by truncating the innermost summation when the variable o is sufficiently large that $\Pr(n,u,t,d,o)$ is negligible. GONOME truncates the sum over o when $\Pr(n,u,t,d,o)$ is decreasing and $\Pr(n,u,t,d,o) < \frac{\varepsilon}{n^2} p$, where p is the current value of the sum (Equation 1) and ε is a user-selected error threshold. It can be shown that $\Pr(n,u,t,d,o)$ decreases monotonically in o once it reaches its maximum, so the total fractional error in the p-value (Equation 1) will be no more than ε.

GONOME saves additional time by only computing Equation 1 if the Z-score of the observed score, S, is greater than a user defined cutoff. The Z-score is computed using the mean and standard deviation of S, which can be computed efficiently. Details of the derivation of the mean and standard deviation of the correlation scoring function, including some extensions, and an analysis showing no significant expectations are lost for a cutoff of three standard deviations, are available at [13].

Extraction of upstream motifs associated to functional groups

Isolation of over-represented upstream motifs is done by extracting all k-mers in the *S. cerevisiae* genome of lengths 5 to 11 that appear five or more times on either strand. For each k-mer, we use GONOME to determine which (if any) GO terms are correlated with its occurrences in the genome. (We restrict the analysis to upstream positions with a 2000 bp cutoff.) Each k-mer is labeled with the GO term with the lowest E-value, provided that the E-value is less than 0.05. The k-mers are then grouped by their GO term labels. Each such group represents a putative motif. The resultant motifs are then filtered to remove those with fewer than four occurrences (of any of the k-mers in the set) in upstream regions and those with more than two hundred occurrences. Motifs with fewer than four occurrences have little statistical support. Motifs with more than two hundred occurrences tend to be such things as TATA motifs, and we chose to ignore them as well.

We then use the MEME [25] algorithm to refine the motifs. Each set of k-mers is input to MEME. MEME aligns the k-mers and creates a position specific scoring matrix (PSSM) for the refined motif. We chose not to use the

PSSM, but instead use the consensus sequence that MEME also outputed as the "final" motif. We then validate each consensus sequence determining all positions in the genome that match the consensus exactly. These positions are treated as the final occurrences of the motif and input to GONOME to see if the original GO term labeling the motif is significantly over-represented. The significant consensus sequences are then compared to known transcription factor consensus motifs.

Datasets

The feature positions used for mouse, human, *D. melanogaster* and *A. thaliana* were extracted from their Genbank genomes. The human genome used herein was NCBI build 35, 26 Aug. 2004, the murine genome was NCBI build 33, 2 Sep. 2004, fly was the 13 Apr. 2005 version, and cress the 19 Feb. 2004 version [26]. The *S. cerevisiae* feature positions were derived from the SGD annotations, 7 Dec. 2004 [27]. GO annotations for human, mouse and *S. cerevisiae* were derived from the 200411 version of the main GO database, (5 Nov. 2004). Feature positions (Chromosome contigs, 4 Jul. 2005) and GO annotation table (15 Aug. 2005) for *S. pombe* came from the Sanger center [28]. *C. elegans* feature positions were derived from the WS147 build GFF file (22 Aug 2005) and GO table revision 1.52 [29]. *D. melanogaster* GO annotations came from the revision 1.65 of the FLYBASE GOA table [30]. *A. thaliana* GO annotations were derived from the TAIR GOA table, revision 1.821 [31]. All datasets are archived at the website [13].

CpG island data was generated using the UCSC version of Larsen's CpG island scanner [11] using default parameters, and taking the position of the 3' end on each strand. The scanner was run over the unmasked human genome, and then those positions matching regions annotated as repetitive by A. F. A. Smit and P. Green's RepeatMasker were removed. The PACE UAS sequences were extracted using S. Weng's PatMatch program at the SGD website [32]. The parameters were set to extract only exact matches.

Transcription factor consensus motifs were drawn from the SGD verified list [33], which was primarily derived from [34].

Availability and requirements

Project name: GONOME

Project homepage: http://gonome.imb.uq.edu.au/.

Operating systems: platform independent

Programming languages: Perl, C++

Other requirements: Optionally Connection to a GO database

License: open source under MIT license

Any restrictions to use by non-academics: No

The downloadable GONOME package includes a BIOPERL [35] based parser for extracting necessary data from Genbank or EMBL files. Input to GONOME consists of a table of the starts, ends, strand and feature IDs of genes, and a list of genomic positions. Output from GONOME is a graph giving the *E*-values of the most over-represented GO terms, and tables providing *E*-values.

The web version of GONOME [6] allows on-line querying against the human, mouse, *S. cerevisiae*, *S. pombe*, *C. elegans*, *D. melanogaster*, and *A. thaliana* (more genomes coming soon) using a set of user-provided genomic positions.

List of abbreviations

EMBL European Molecular Biology Laboratory

GFF General Feature Format

GO Gene Ontology

NCBI National Center for Biotechnology Information

PACE Proteosome Associated Control Element

PSSM Position Specific Score Matrix

UAS Upstream Activating Sequence

UCSC University of California, Santa Cruz.

Authors' contributions

SMS: Conceived and developed the application, implemented the analyses and drafted the manuscript.

TLB: Proposed much of the statistical methodology, advised on usage of MEME application and assisted in drafting the manuscript.

JSM: Advised, participated in interface design and assisted in drafting the manuscript.

All authors read and approved the final manuscript.

Acknowledgements

We thank Geoff McLachlan and Michael Gagen for helpful discussions and useful feedback on this work.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
2. Beissbarth T, Speed TP: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20:**1464-1465.
3. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO::TermFinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.** *Bioinformatics* 2004.
4. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304:**1321-1325.
5. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15:**1034-1050.
6. **GONOME - Finding associations between genomic positions and Gene Ontology (GO) terms** [http://gonome.imb.uq.edu.au/]
7. Robinson PN, Bohme U, Lopez R, Mundlos S, Nurnberg P: **Gene-Ontology analysis reveals association of tissue-specific 5' CpG-island genes with development and embryogenesis.** *Hum Mol Genet* 2004, **13:**1969-1978.
8. Antequera F: **Structure, function and evolution of CpG island promoters.** *Cell Mol Life Sci* 2003, **60:**1647-1658.
9. Ponger L, Duret L, Mouchiroud D: **Determinants of CpG islands: expression in early embryo and isochore structure.** *Genome Res* 2001, **11:**1854-1860.
10. Antequera F, Bird A: **Number of CpG islands and genes in human and mouse.** *Proc Natl Acad Sci U S A* 1993, **90:**11995-11999.
11. Larsen F, Gundersen G, Lopez R, Prydz H: **CpG islands as gene markers in the human genome.** *Genomics* 1992, **13:**1095-1107.
12. Scherer SW, Cheung J, MacDonald JR, Osborne LR, Nakabayashi K, Herbrick JA, Carson AR, Parker-Katiraee L, Skaug J, Khaja R, Zhang J, Hudek AK, Li M, Haddad M, Duggan GE, Fernandez BA, Kanematsu E, Gentles S, Christopoulos CC, Choufani S, Kwasnicka D, Zheng XH, Lai Z, Nusskern D, Zhang Q, Gu Z, Lu F, Zeesman S, Nowaczyk MJ, Teshima I, Chitayat D, Shuman C, Weksberg R, Zackai EH, Grebe TA, Cox SR, Kirkpatrick SJ, Rahman N, Friedman JM, Heng HH, Pelicci PG, Lo-Coco F, Belloni E, Shaffer LG, Pober B, Morton CC, Gusella JF, Bruns GA, Korf BR, Quade BJ, Ligon AH, Ferguson H, Higgins AW, Leach NT, Herrick SR, Lemyre E, Farra CG, Kim HG, Summers AM, Gripp KW, Roberts W, Szatmari P, Winsor EJ, Grzeschik KH, Teebi A, Minassian BA, Kere J, Armengol L, Pujana MA, Estivill X, Wilson MD, Koop BF, Tosi S, Moore GE, Boright AP, Zlotorynski E, Kerem B, Kroisel PM, Petek E, Oscier DG, Mould SJ, Dohner H, Dohner K, Rommens JM, Vincent JB, Venter JC, Li PW, Mural RJ, Adams MD, Tsui LC: **Human chromosome 7: DNA sequence and biology.** *Science* 2003, **300:**767-772.
13. **GONOME resources and examples** [http://gonome.imb.uq.edu.au/Resources.html]
14. Kapranov AB, Kuriatova MV, Preobrazhenskaia OV, Tiutiaeva VV, Shtuka R, Feldmann H, Karpov VL: **[Isolation and identification of PACE-binding protein rpn4--a new transcription activator, participating in regulation of 26S proteosome and other genes].** *Mol Biol (Mosk)* 2001, **35:**420-431.
15. Cora D, Di Cunto F, Provero P, Silengo L, Caselle M: **Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs.** *BMC Bioinformatics* 2004, **5:**57.
16. Mannhaupt G, Schnall R, Karpov V, Vetter I, Feldmann H: **Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast.** *FEBS Lett* 1999, **450:**27-34.
17. Wang L, Mao X, Ju D, Xie Y: **Rpn4 is a physiological substrate of the Ubr2 ubiquitin ligase.** *J Biol Chem* 2004, **279:**55218-55223.
18. Ashe MP, De Long SK, Sachs AB: **Glucose depletion rapidly inhibits translation initiation in yeast.** *Mol Biol Cell* 2000, **11:**833-848.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
20. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I: **Controlling the false discovery rate in behavior genetics research.** *Behav Brain Res* 2001, **125:**279-284.
21. Dalmasso C, Broet P, Moreau T: **A simple procedure for estimating the false discovery rate.** *Bioinformatics* 2004.
22. Storey JD: **A direct approach to false discovery rates.** *J Royal Statistical Soc B* 2002, **64:**479-498.
23. Manly KF, Nettleton D, Hwang JTG: **Genomics, Prior Probability, and Statistical Tests of Multiple Hypotheses.** *Genome Res* 2004, **14:**997-1001.
24. Bishop YMM, Feinberg SE, Holland PW: **Discrete multivariate analysis : theory and practice.** Cambridge,Massachusetts and London,England, The MIT Press; 1975.
25. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**28-36.
26. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucl Acids Res* 2005, **33:**D501-504.
27. **Saccharomyces cerevisiae data** [ftp://genome-ftp.stanford.edu/pub/yeast/chromosomal_feature/SGD_features.tab]
28. **Schizosaccharomyces pombe data** [ftp://ftp.sanger.ac.uk/pub/yeast/pombe/]
29. **Caenorhabditis elegans data** [ftp://ftp.wormbase.org/pub/wormbase/genomes/elegans]
30. **Drosophila melanogaster data** [ftp://flybase.bio.indiana.edu/genomes/Drosophila_melanogaster/]
31. **The Arabidopsis Information Resource** [http://www.arabidopsis.org]
32. **SGD Help: Pattern Matching** [http://www.yeastgenome.org/help/nph-patmatch.html]
33. **Transcription factor consensus motifs** [ftp://genome-ftp.stanford.edu/pub/yeast/data_download/systematic_resultregulatory_regions/harbison_pmid_15343339/Motifreferences-veri fied.txt]
34. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431:**99-104.
35. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.